VQEG eLetter

Volume 1, Issue 1 • October 2014

Dreamed about training, verifying and validating your QoE model on a million videos?

Glenn Van Wallendael, Nicolas Staelens, Enrico Masala, Lucjan Janowski, Kongfeng Berger, Marcus Barkowsky

Introduction

Although we are not yet at a million videos, gradual additions over time will eventually get us there. In the beginning of the large-scale database effort, in 2012, the main focus was on encoding conditions.

Training, verification, and validation of objective prediction models require well-chosen test stimuli. The measured prediction performance depends largely on the congruence of stimulus selection in the three steps training, verification, and validation. Different stimulus selection criteria are controversially discussed: extracting a representative set of stimuli from the scope of application, spreading the range of application scope with equidistant stimuli, or using stressful stimuli for the prediction algorithm. Nowadays, most databases are too small to sufficiently cover even one of these evaluation types, a large-scale database may solve the problem but requires new statistical methods and understanding of quality evaluation.

Therefore, it all started with 10 HD-sequences which were also downscaled by a factor of 4 and 8. They were encoded with 430 different encoding parameters like bitrate, frame rate, encoding structure, encoder implementation, number of slices, and so on, resulting in 12960 H.264/AVC encoded video streams. These sequences were annotated by Full-Reference (FR) results. The same video sequence set got encoded with the H.265/HEVC standard as well employing 5952 different encoding settings leading to another set of 59520 encoded sequences.

What's the quality of each of these sequences? While a full subjective experiment is prohibitive, objective algorithms may be computed and compared, stimulating research on new types of agreement analysis. Currently, the database features 5 video quality metrics computed for each encoded video sequence, which are: Peak Signal to Noise Ratio (PSNR)[1], Structural Similarity Index (SSIM)[2], Visual Information Fidelity (VIF)[3], Video Quality Metric (VQM)[4], and Perceptual Video Quality Measure (PVQM)[5]. Further details are available on the JEG wiki[6].

Efforts are under way to extend the database in the direction of adding more content, notably also in Ultra-HD resolution, as well as to provide the same measures for sequences impaired by packet losses. To this aim, an H.265/HEVC robust decoder[7] has been used to produce distorted video sequences on the basis of 25 different loss patterns. Although it is difficult to provide such measures for all loss patterns applied to all the encoded sequences due to the huge processing time required, it is expected that in the next six months at least a significant subset of the original encoded video sequences will have all the quality measures corresponding to the 25 loss patterns.

---

[1] NTIA / ITS. (2001). A3: Objective Video Quality Measurement Using a Peak-Signal-to-Noise-Ratio (PSNR) Full Reference Technique. ATIS T1.TR.PP.74-2001

[2] NTIA / ITS. (2001). A3: Objective Video Quality Measurement Using a Peak-Signal-to-Noise-Ratio (PSNR) Full Reference Technique. ATIS T1.TR.PP.74-2001

[3] Sheikh, H. R., &Bovik, A. C. (2006).Image information and visual quality. IEEE Transactions on Image Processing, 15(2), 430–444.

[4] ITU-T Study Group 9.(2004). ITU-T J.144 Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference. ITU-T J.144

[5] Hekstra, A. P., Beerends, J. G., Ledermann, D., de Caluwe, F. E., Kohler, S., Koenen, R. H., et al. (2002). PVQM – A perceptual video quality measure. Elsevier, Signal Processing: Image Communications 17, , 781–798.

[6] http://vqegjeg.intec.ugent.be/wiki/index.php/JEG_no-reference_hybrid_HEVC

[7] http://media.polito.it/jeg

Development and performance evaluations of objective assessment algorithms

Most industrial and research effort has been spent so far on creating holistic objective assessment algorithms optimized for a particular application scenario. Rarely the intermediate steps of such complex algorithms have been evaluated separately.

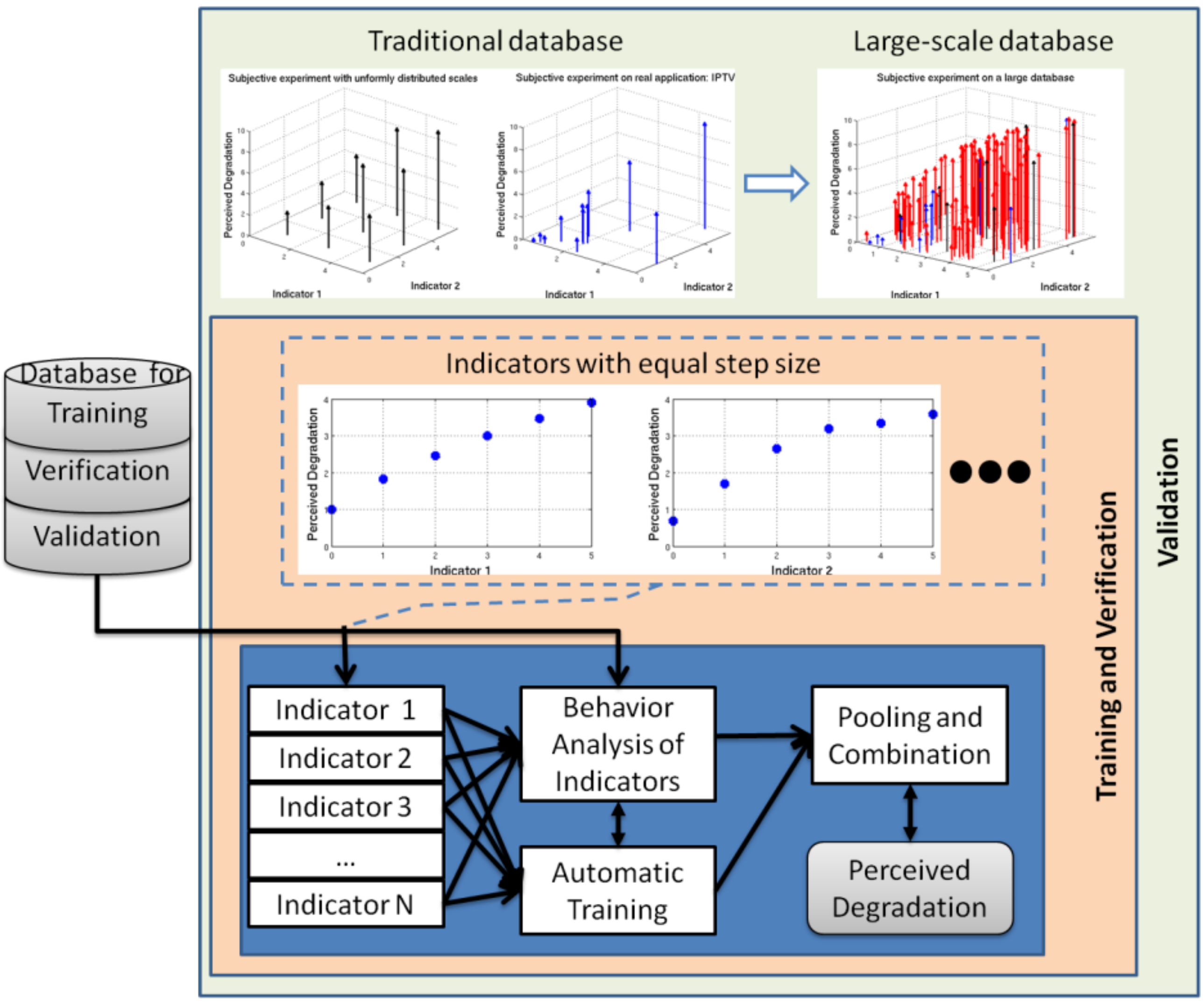

Figure 1: An overview of a typical development cycle of objective quality assessment

Figure 1 shows a functional overview of the typical development cycle. The cycle, in general, includes a training procedure followed by verification, and after development has finished, validation is performed. In the training procedure, various indicators are developed, pooled over space and time, and then merged to predict the perceived quality. Typical prediction performance measures include linearity (Pearson Linear Correlation Coefficient, PLCC), Rank Ordering (Spearman Rank Order Coefficient, SROCC), and accuracy (Root Mean Square Error, RMSE). The stability of the estimated fitting parameter during training and the appropriateness of its count as compared to the samples available for training may be evaluated by cross-validation of the training process.

Validation requires a different set of samples. In the validation procedure, algorithms of objective quality assessment are often validated using the same performance measures as previously introduced for verification. In addition, more sophisticated measure may be used, for example Epsilon Insensitive RMSE (RMSE*), Outlier Ratio with respect to Standard Error as detailed in ITU-T P.1401, and Accuracy Analysis or Resolving Power as specified by ITU-T J.149.

A typical objective video quality assessment algorithm combines several quality indicators where each of them should ideally provide good quality prediction results when used within its scope of application, rough estimates when used at the boundaries or in an extended scope and each of them should stay neutral when confronted with degradations out of its specific measurement scope. A typical example would be a perceptual frame rate indicator that correctly predicts constant frame rate settings, that has limited accuracy when the frame rate becomes variable, and that stays neutral when longer pauses and skips occur as those isolated events require a different perceptual measurement[8].

Figure 1 shows the systematic development situation of a quality prediction algorithm in a block diagram. Several perceptual features are identified and experimented in isolated subjective experiments such that the degradations occur equally often in different strengths. The expected behavior of each indicator with respect to subjective results is illustrated by the two plots in the orange verification procedure block. This process may be simplified as a one dimensional training procedure for each indicator algorithm but in practice the indicators are interdependent. For example, the ratio of frame rate reduction is dependent on resolution in the application scenario of IPTV.

How is a large database going to help in the development stage?

Most objective metrics were designed for certain applications, such as compression only[9], or compression and transmission degradations, additionally including display post-processing and so on. The existing databases were also built for certain applications. Metrics developed for compression may perform well on the database of compressed videos, and it is very likely that these metrics were tested only on compressed videos. It is of great interest to know how these distortion-specific metrics perform on videos in their extended scope or out of their scope. For example, how a metric designed for H.264 compressed natural videos performs on HEVC compressed videos, videos with packet loss, and computer-generated videos. Observing the performance of distortion-

[8]Barkowsky, Staelens, Janowski, Koudota, Leszczuk, Urvoy, et al. (2012). Subjective experiment dataset for joint development of hybrid video quality measurement algorithms .QoEMCS 2012, Berlin

[9]K. Zhu, C. Li, V. K. Asari, and D. Saupe, "No-reference video quality assessment based on artifacts measurement and statistical analysis." IEEE Transactions on Circuits and Systems for Video Technology, 2014.

specific metrics on videos in their extended scope and out of their scope calls for a large-scale database with videos impacted by various degradations.

Another problem that may be solved by a large database is machine-learning based algorithms' over-fitting. Machine-learning based algorithms, in general, have good quality prediction accuracy. They are, however, highly prone to over-fitting on the training set, and therefore end up with a low generalizability[10]. In many cases, the number of videos in the training set is small in comparison to the large number of parameters in the trained algorithm. Additionally, the content of videos in the training set is not diverse enough. Consequentially, the predicted quality of the model may show large errors to the MOS when a video has different content from the training videos. Both problems, over-fitting and lack of considered content, can be avoided by large databases. Typically, machine learning methods' stability are evaluated by cross-validation. For example, the 10-fold cross-validation is an often used strategy to assess how a machine-learning based algorithm performs on unseen data. We noticed that the statistical results of cross-validation are sensitive to cross-validation strategy and the number of video sets in one fold. With a large video database, the number of video sets in one fold is also large such that the cross-validation results are robust, and therefore, the estimated general performance of machine-learning based algorithm on unseen data is robust.

How is a large database going to improve the validation stage?

Performance evaluation with respect to the application scenario is the primary purpose of the validation step. Previous VQEG efforts on SDTV, Multimedia, HDTV, and

[10]P. Gastaldo and J. A. Redi, "Machine learning solutions for objective visual quality assessment," in the sixth International Workshop on Video Processing and Quality Metrics, Jan. 2012.

Hybrid models[11] document the enormous effort required for this black box type of independent validation of computational models.

The selection of both the source content (SRC) and the degradation, also called Hypothetical Reference Circuit (HRC) forms a crucial part of such evaluation.. Open questions include whether the coverage of samples shall be uniform with respect to the scope of application, i.e. as many perfect as average as strongly degraded videos, or uniform with respect to the expected application scenario, i.e. more average quality videos than perfect or strongly degraded videos. Figure 1 shows this graphically in the green validation area. The first two diagrams illustrate the situation in case that the validation database is designed for equally covering the scope of the indicators which may or may not coincide with covering equally the application scope.

The second diagram illustrates the distribution when focusing on typical application examples: In most real-world services, the perceived quality is above average most of the time and strong degradations occur rather seldom. The third diagram illustrates that a large-scale database allows for both types of evaluations and actually may inverse the interpretation: It may provide the answer on which application scopes an algorithm is applicable to besides the one that it was designed for.

This question also applies to content. The choice of extreme contents, such as artistic video sequences, may bias the evaluation while allowing for the analysis of the stability of the algorithms. A large-scale database would therefore allow for more detailed analysis including overall suitability of quality prediction algorithms and their behavior at the limits of the application scope.

---

[11] VQEG reports from completed validation tests, online: http://www.its.bldrdoc.gov/vqeg/reports.aspx

Glenn Van Wallendael *obtained the M.Sc. degree in Applied Engineering from the University College of Antwerp, Belgium, in 2006 and the M.Sc. degree in Engineering from Ghent University, Belgium in 2008. Afterwards, he worked towards a Ph.D. at Multimedia Lab, Ghent University, with the financial support of the Agency for Innovation by Science and Technology (IWT). Currently, he continues working in the same group as a post-doctoral researcher. His main topics of interest are video compression including scalable video compression and transcoding.*

Nicolas Staelens *obtained his Master's degree in Computer Science at Ghent University (Belgium, 2004).In 2006, he joined the Internet Based Communication Networks and Services (IBCN) group at Ghent University where he received a Ph.D. degree in Computer Science Engineering in February 2013. The topic of his dissertation was "Objective and Subjective Quality Assessment of Video Distributed over IP-based Networks".As of 2007, he is also actively participating within the Video Quality Experts Group (VQEG) and is currently co-chair of the Tools and Subjective Labs support group and the JEG-Hybrid project.*

More detailed analysis may also be obtained with respect to the accuracy of an indicator measuring a technical parameter (i.e. bitrate), a perceptual feature (i.e. blockiness), or a complete algorithm within a certain quality range, i.e. near-lossless or strongly degraded. The combination of several algorithms may be proposed during validation[12].

The availability of a variety of SRC and HRC used for validation is often a bottleneck in traditional approaches.

A large-scale approach may have such a large selection of both SRC and HRC that conducting a formal subjective assessment on a subset may be considered sufficient for validation. Otherwise, the reproducible processing for the creation of the database may simplify the creation of similar or completely new processed sequences. Evaluating algorithms on each result obtained in the large-scale database allow for drawing a complete picture of its stability ,applicability to a certain (sub-)scope, comparing with other available algorithms. An example would be to provide a resolving power analysis for each application that may be automatically predicted in a next step.

Sample results

To give a rough idea of the possibilities opened by the currently available large-scale database[13], a sample validation result is reported here. When taking any two video sequences from the large scale data set and evaluating their quality with either PSNR, SSIM, or VIF, a rank order can be established. It would be interesting to understand to what extent the three measures agree on the ranking. For three measures, there will

[12]Barri, A.; Dooms, A.; Jansen, B.; Schelkens, P., "A Locally Adaptive System for the Fusion of Objective Quality Measures," Image Processing, IEEE Transactions on , vol.23, no.6, pp.2446,2458, June 2014

[13]M. Barkowsky, E. Masala, G. Van Wallendael, K. Brunnström, N. Staelens, P. Le Callet, "Objective Video Quality Assessment – Towards large scale video database enhanced model development", IEICE Transactions Quality of Diversifying Communication Networks and Services, 2015, accepted for publication

__Enrico Masala__ received the Ph.D. degree in computer engineering from the Politecnico di Torino, Turin, Italy, in 2004. In 2003, he was a visiting researcher at the Signal Compression Laboratory, University of California, Santa Barbara, where he worked on joint source channel coding algorithms for video transmission. Since 2011 he is Assistant Professor in the Control and Computer Engineering Department at the Politecnico di Torino. His main research interests include simulation and performance optimization of multimedia communications (especially video) over wireline and wireless packet networks.

__LucjanJanowski__ is an assistant professor at the Department of Telecommunications, AGH University of Science and Technology, in Krakow, Poland. He is a Co-Chair of the VQEG JEG-Hybrid project (http://www.its.bldrdoc.gov/vqeg/projects/jeg/jeg.aspx).

be either agreement or exactly one metric which does not agree.

For each measure we calculate the distance between the two sequences in a pair when the measure disagrees. There is a total of six possible cases, i.e., for each one of the three measures, one of the other two does not agree.

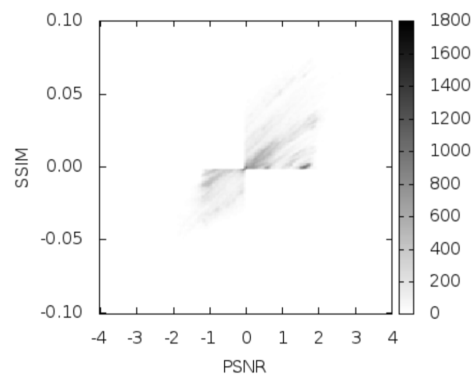

Figure 2: Density plot of the difference of SSIM and PSNR in the pairwise comparison when VIF disagrees

The scatter plot in Figure 2 represents all pairs of encoded video sequences for src06 when VIF disagrees with PSNR and SSIM. The grey level represents the number of sequences that do not agree for a certain difference of the PSNR and SSIM on the x and y axes, the darker, the more measures disagree. It can be seen that beyond a certain difference in each measure the quality difference is so pronounced that all metrics agree. This limit is approximately +-2 dB for PSNR and +-0.05 for SSIM on their natural scales.

Selecting the 95 percentile value, a reasonable threshold for the prediction consistency of the measure with respect to the two others may be determined. As can be seen from Figure 3, this value is strongly sequence dependent (compare, for instance, sequence 01 and sequence 03 for PSNR), and within the same sequence, there can be a large difference depending on the cause of disagreement (see, e.g., sequence 08).

This shows the advantage of having a large set of coding conditions for measuring the influence of content on a quality measure in validation. Please note that this analysis is purely

Kongfeng Berger *was born in Shandong, China. She received the M.Sc. degree in communication and information systems from Shandong University, China, in 2006, and the Ph.D. degree in multimedia signal processing from the University of Konstanz, Germany, in 2014. She is currently aa Postdoctoral Fellow at the University of Nantes, France. Her main research interests include visual quality assessment, image processing motion analysis, natural scene statistics, feature selection, and machine learning.*

Marcus Barkowsky *received the Dr.-Ing. degree from the University of Erlangen-Nuremberg in 2009. He joined the Image and Video Communications Group at IRCCyN at the University of Nantes in 2008, and was promoted to associate professor in 2010. His activities range from modeling effects of the human visual system, in particular the influence of coding, transmission, and display artifacts in 2D and 3D to measuring and quantifying visual discomfort and visual fatigue on 3D displays using psychometric and medical measurements. He currently co-chairs the VQEG "3DTV"and "Joint Effort Group Hybrid" activities.*

based on disagreement, subjective experiments are required to determine whether the disagreement of one measure with respect to the two others indicate a failure of that measure and whether an agreement of the three measures is consistent with human observations.

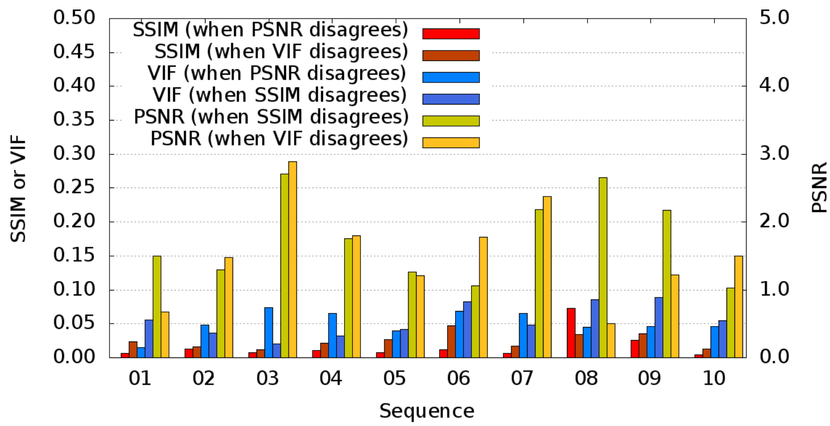

Figure 3: 95 percentile of the two agreeing video quality measures when one disagrees

What's next?

Establishing large-scale databases is a continuous effort, packet losses and higher resolutions as well as more content and encoders need to be added for improving the training, verification and validation process. Further statistical analysis tools should be researched in parallel, innovative analysis questions may be invented as shown with the example above.